

O Teste do Sinal para Amostras Associadas

Carla Henriques

CMUC e Escola Sup. Tecnologia, Inst. Polit. de Viseu - carlahenriq@estv.ipv.pt

Paulo Eduardo Oliveira

CMUC e Universidade de Coimbra - paulo@mat.uc.pt

Resumo: Em Dewan e Prakasa Rao (2005) é estabelecida a normalidade assintótica da estatística do teste do sinal para amostras associadas. Contudo, este resultado depende da variância da distribuição limite, σ^2 , que em geral não é conhecida, pois incorpora as distribuições conjuntas das variáveis de base. Recorrendo aos resultados de Henriques e Oliveira (2006, 2008), estima-se σ e obtém-se a normalidade assintótica da estatística que recorre à estimação de σ . Neste trabalho apresentam-se alguns resultados de um estudo de simulação levado a cabo com o objectivo de avaliar a aproximação à Normal das estatísticas de teste precedentes. Serão analisadas as caudas e a distribuição global das estatísticas de teste e é ilustrado o comportamento do estimador da variância σ^2 .

Palavras-chave: Teste do sinal, normalidade assintótica, simulação.

Abstract: Dewan e Prakasa Rao (2005) establish the asymptotic normality of the test statistic for the signal test when the observed variables are positive associated. However, this result depends on the asymptotic variance, σ^2 , which is usually not known as it depends on the joint distributions of the variables. Using the results of Henriques e Oliveira (2006, 2008), we can estimate σ and obtain the asymptotic normality of the test statistic which involves the estimator of σ . To access the quality of the normal approximation of those test statistics, we present here the results of a simulation study, in which we analyze the tails and the global distribution of the test statistics, and also illustrate the behavior of the estimator of σ .

Keywords: Signal test, asymptotic normality, simulation.

1 Introdução, Pressupostos e Resultados Fundamentais

No presente trabalho estudamos o comportamento do teste do sinal baseado em amostras positivamente associadas, noção que recordamos em seguida.

Definição 1.1 *Dada uma sucessão de variáveis aleatórias $\{X_n, n \geq 1\}$, dizemos que a sucessão é associada, se para todo o $n \in \mathbb{N}$ e para todo o par de funções f e g de \mathbb{R}^n em \mathbb{R} não decrescentes em cada variável,*

$$\text{Cov}(f(X_1, \dots, X_n), g(X_1, \dots, X_n)) \geq 0,$$

sempre que esta covariância exista.

Para uma exposição de resultados e conceitos fundamentais sobre associação, referimos os artigos de Newman (1984), Suquet (1994), Roussas (1999) e Dewan e Prakasa Rao (2001), que incluem alguns dos resultados mais relevantes obtidos no contexto da associação, bem como inúmeras referências e diversos exemplos de aplicação deste conceito.

O teste do sinal baseado num amostra associada é estudado pela primeira vez em Dewan e Prakasa Rao (2005), onde se deduz a normalidade assintótica da estatística de teste. Contudo, a estatística utilizada para este teste depende da variância da distribuição limite, σ^2 , que, em geral, não é conhecida, pois depende de uma série que envolve as distribuições conjuntas de (X_1, X_{k+1}) , $k = 1, 2, \dots$. Os trabalhos de Henriques e Oliveira (2006, 2008) propiciam uma forma de contornar esta limitação, na medida em que neles é estudado um estimador para esta série, sendo obtidos resultados de consistência deste estimador, com caracterização de velocidades de convergência.

Antes de apresentar o resultado de Dewan e Prakasa Rao (2005), indicamos de seguida as condições que serão utilizadas nos resultados incluídos neste artigo. No que se segue, representa-se por $F(s)$ a função de distribuição de X_1 e por $F_k(s, t)$ a função de distribuição do par (X_1, X_{k+1}) . A primeira condição exprime a dependência entre as variáveis.

(S) As variáveis aleatórias $\{X_n, n \geq 1\}$ constituem uma sucessão associada e estritamente estacionária de variáveis aleatórias com função densidade limitada por uma constante B_0 .

Relativamente à estrutura de covariâncias de $\{X_n, n \geq 1\}$, definindo, $u(n) = \sum_{j=n+1}^{\infty} \text{Cov}^{1/3}(X_1, X_j)$, $n \geq 1$, serão consideradas as seguintes condições:

(C1) $u(0) < +\infty$;

(C2) existem $\theta > 0$ e $C_\theta > 0$ tal que $u(n) \leq C_\theta n^{-\theta}$, para todo o $n \geq 1$.

Serão ainda consideradas condições explícitas no decrescimento das covariâncias $\text{Cov}(X_1, X_n)$:

(P) $\text{Cov}(X_1, X_n) = a_0 n^{-a}$, para algum $a_0 > 0$ e algum $a > 3$;

(G) $\text{Cov}(X_1, X_n) = a_0 a^{-n}$, para algum $a_0 > 0$ e algum $a > 1$.

Refira-se que condições deste tipo são comuns na literatura sobre associação.

Denotemos por M_e a mediana da distribuição de X_1 . Para testar a hipótese nula $H_0 : M_e = \theta$, onde θ é um valor fixo à partida, o teste do sinal baseia-se, como é sabido, na estatística $U_n = \frac{1}{n} \sum_{i=1}^n I_{(\theta, +\infty)}(X_i)$.

Teorema 1.2 (Dewan e Prakasa Rao, 2005) *Seja $\{X_n, n \geq 1\}$ uma sucessão de variáveis aleatórias que satisfaz (S) e (C1). Defina-se*

$$\sigma^2 = F(\theta) - F(\theta)^2 + 2 \sum_{k=1}^{\infty} [F_k(\theta, \theta) - F(\theta)^2] .$$

Tem-se então

$$T_{\sigma,n} = \frac{\sqrt{n}(U_n - P(X_1 > \theta))}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

Este teorema sugere que, para o teste da hipótese $H_0 : M_e = \theta$ contra a alternativa $H_1 : M_e > \theta$, o procedimento de decisão consista em rejeitar a hipótese nula H_0 para valores grandes da estatística U_n .

O teorema anterior é na verdade um caso particular do resultado estabelecido em Bagai e Prakasa Rao (1991), relativo à convergência das distribuições de dimensão finita da função de distribuição empírica unidimensional, assim como do resultado obtido em Henriques e Oliveira (2003) que generaliza aquele para a função de distribuição empírica bidimensional.

Como já se referiu anteriormente, o procedimento de decisão definido no teorema anterior só se poderá pôr em prática se σ for conhecido, situação que em geral se não verifica. Esta limitação pode ser ultrapassada recorrendo aos resultados de Henriques e Oliveira (2006, 2008) que estabelecem condições suficientes para a consistência de um estimador para a série $\Psi(s, t) = \sum_{k=1}^{\infty} [F_k(s, t) - F(s)F(t)]$.

A estimação das funções $F(s)$ e $F_k(s, t)$ é feita recorrendo às funções de distribuição empírica uni e bidimensional, respectivamente, definidas por $\hat{F}_n(s) = \frac{1}{n} \sum_{i=1}^n (I_{(-\infty, s]}(X_i))$ e $\hat{F}_{k,n}(s, t) = \frac{1}{n-k} \sum_{i=1}^{n-k} (I_{(-\infty, s]}(X_i)I_{(-\infty, t]}(X_{i+k}))$.

O estimador para a série em causa será então definido por

$$\hat{\Psi}_n(s, t) = \sum_{k=1}^{q_n} [\hat{F}_{k,n}(s, t) - \hat{F}_n(s)\hat{F}_n(t)],$$

onde q_n determina o número de termos a somar para aproximar a série e deve ser tal que $q_n \rightarrow +\infty$ e $q_n/n \rightarrow 0$.

Nos trabalhos de Henriques e Oliveira (2006, 2008) são apresentadas condições suficientes para a consistência uniforme forte deste estimador, identificando-se velocidades de convergência. Em resumo, q_n deve ter um crescimento controlado para se atingir a melhor velocidade de convergência do estimador.

Ora, sob H_0 , tem-se $\sigma^2 = 1/2 - 1/4 + 2 \sum_{k=1}^{\infty} [F_k(\theta, \theta) - F(\theta)^2]$, quantidade que pode ser estimada por

$$\hat{S}_n^2 = \frac{1}{4} + 2 \sum_{j=1}^{q_n} [\hat{F}_{j,n}(\theta, \theta) - \hat{F}_n(\theta)^2].$$

Nas condições dos Teoremas 3.3 ou 3.4 de Henriques e Oliveira (2006) ou, ainda, do Lema 4.1 ou Teorema 4.2 de Henriques e Oliveira (2008), tem-se então a consistência de \hat{S}_n^2 . Aplicando o Teorema de Slutsky, segue-se o seguinte teorema, que resume as condições dos resultados de Henriques e Oliveira (2006, 2008).

Teorema 1.3 *Seja $\{X_n, n \geq 1\}$ uma sucessão de variáveis aleatórias que verifica (S). Se $M_e = \theta$, então*

$$T_{\hat{S},n} = \frac{\sqrt{n}(U_n - 1/2)}{\hat{S}_n} \xrightarrow{d} \mathcal{N}(0, 1),$$

desde que se verifique uma das condições seguintes:

- (a) $\{X_n, n \geq 1\}$ satisfaz (G) e $q_n = O(n^{1/3} \log^{-\gamma} n)$, para algum $\gamma > 2/3$;
- (b) $\{X_n, n \geq 1\}$ satisfaz (C2) com $\theta = \frac{p-2}{2}$, para algum $p > 2$, e $q_n \sim n^{\frac{p-2-2\delta}{p^2+3p}}$, para algum $0 < \delta < \frac{p-2}{2}$;
- (c) $\{X_n, n \geq 1\}$ satisfaz (G) e $q_n \sim n^{\frac{p-2-2\delta}{p^2+3p}}$, para algum $p > 2$ e algum $0 < \delta < \frac{p-2}{2}$;
- (d) $\{X_n, n \geq 1\}$ satisfaz (P) e $q_n \sim n^{\frac{3}{a-3}\gamma}$, para algum $0 < \gamma < \frac{1}{2} - \frac{21a-18}{2a(2a+9)}$.

2 Estudo de Simulação

Com o intuito de avaliar a aproximação à normal para amostras finitas das distribuições das estatísticas de teste $T_{\sigma,n}$ e $T_{\hat{S},n}$, foram geradas amostras associadas satisfazendo todas as condições dos resultados da secção anterior. Estas amostras foram obtidas recorrendo a três algoritmos diferentes que passamos a descrever. Constrói-se uma sequência Y_1, Y_2, \dots , de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) e aplica-se um dos seguintes procedimentos:

1. $Y_i \sim \mathcal{N}(0, 1)$, fixa-se $m \in \mathbb{N}$ e toma-se $X_i = (m+1)^{-1/2}(Y_i + \dots + Y_{i+m})$;
2. $Y_i \sim \exp(1)$, fixa-se $m \in \mathbb{N}$ e toma-se $X_i = (m+1) \min(Y_i, \dots, Y_{i+m})$;
3. $Y_i \sim \mathcal{N}(0, 1)$, fixa-se $\theta \in (0, 1)$ e toma-se $X_i = Y_i + \theta X_{i-1}$.

A sucessão X_1, X_2, \dots , construída por qualquer um dos algoritmos precedentes, satisfaz as condições (S) e (G), logo também (C1) e (C2). Além disso, é possível determinar a variância σ^2 , o que permite não só avaliar o comportamento dos estimador \hat{S}_n^2 , como também comparar as estatísticas $T_{\sigma,n}$ e $T_{\hat{S},n}$. Os valores pré-fixados de m e de θ estão directamente relacionados com o grau de dependência das variáveis da sucessão construída. Note-se, ainda, que os dois primeiros algoritmos geram sucessões m -dependentes, o que significa que X_i e X_{i+k} são independentes para $k > m$. Consequentemente, a série $\Psi(s, t)$ terá apenas m termos não nulos. O mesmo não se passa com o Algoritmo 3, pois este gera uma média móvel infinita.

O Teorema 1.3 diz-nos qual deve ser a velocidade de crescimento da sucessão q_n , mas não dispomos de qualquer informação sobre eventuais constantes multiplicativas. Obviamente, para amostras finitas a informação sobre o crescimento da sucessão q_n não é suficiente. De facto, para um n fixo, será muito diferente

tomar, por exemplo, $q_n = 10n^\beta$ ou $q_n = 1000n^\beta$, para um qualquer $\beta > 0$, embora, em ambos os casos a velocidade de crescimento de q_n seja a mesma.

Pretendendo-se comparar globalmente as distribuições das estatísticas $T_{\sigma,n}$ e $T_{\hat{\sigma},n}$, construíram-se os dois gráficos apresentados na Figura 1. Além destas duas estatísticas, considerou-se também, para efeitos comparativos, a estatística do teste do sinal obtida para amostras de observações independentes de lei $N(0,1)$, que denotaremos simplesmente por T_n . Em cada um dos gráficos são pois apresentadas quatro curvas, a curva da função densidade da normal padrão, que serve de referência, juntamente com mais três curvas, uma para cada uma das três estatísticas referidas ($T_{\hat{\sigma},n}$, $T_{\sigma,n}$, T_n). As três últimas curvas foram obtidas pelo método do núcleo. O algoritmo usado para gerar as amostras associadas foi o Algoritmo 3, com $\theta = 0.2$ no caso do gráfico à esquerda e $\theta = 0.8$, o que reflecte um maior grau de dependência, no caso do gráfico à direita.

Podemos pois constatar, por análise dos gráficos da Figura 1, que o comportamento da estatística $T_{\sigma,n}$ é muito semelhante ao da estatística T_n , referente a amostras de observações independentes, mesmo quando o grau de dependência é maior (gráfico à direita). Quanto ao estimador $T_{\hat{\sigma},n}$, este apresenta uma pior aproximação à normal, que se agrava quando o grau de dependência aumenta.

Se em vez do Algoritmo 3 tivéssemos usado um dos outros dois algoritmos, iríamos observar o mesmo tipo de comportamento das três estatísticas.

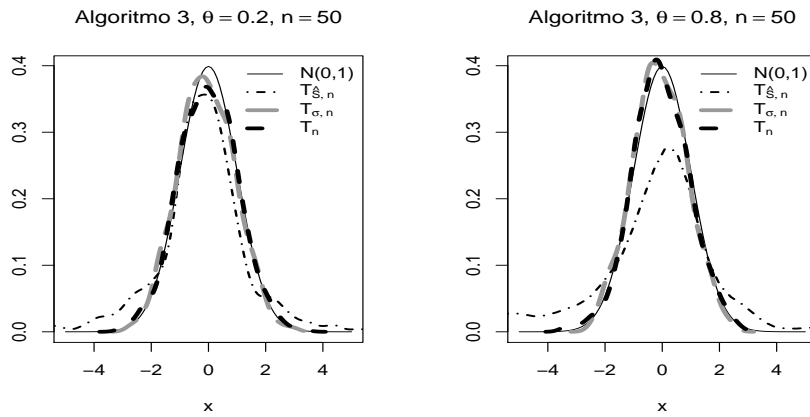


Figura 1: Comparação das estatísticas $T_{\hat{\sigma},n}$, $T_{\sigma,n}$ e T_n .

O estudo de simulação mostrou-nos também que o número de parcelas ($q_n \equiv q$), que tomamos para aproximar a série na estimação de σ , afecta bastante os resultados. Na verdade, como já foi dito anteriormente, quando as sucessões

são geradas pelos Algoritmos 1 e 2, a série $\Psi(s, t)$ terá apenas m termos não nulos. Nestes casos, como é de esperar, os melhores resultados são obtidos quando somamos $q = m$ termos para aproximar a série. Além disso, constatámos também que entre tomar para q um valor grande (superior a m) ou um valor pequeno (inferior a m) é preferível a primeira situação, principalmente se a dimensão amostral for bastante grande ($n > 100$). No caso do Algoritmo 3, apesar dos termos da série nunca se anularem, a partir de uma certa ordem estes são tão pequenos que tudo se passa como nos Algoritmos 1 e 2.

Na verdade o comportamento que acabámos de descrever parece-nos natural. De facto, sendo $\Psi(s, t)$ uma série convergente, os seus termos decrescem para zero. Contudo, quando tomamos para q um valor muito pequeno, estamos a sub-estimar a série, obtendo-se portanto resultados pobres. Valores de q grandes serão portanto preferíveis a valores de q pequenos. Mesmo na situação em que a série tem na verdade apenas m parcelas não nulas e tomamos $q > m$, as estimativas da soma não serão más, porque os termos que somamos a mais estão a estimar parcelas iguais a zero, sendo por isso numericamente pequenos. Claramente, esta sobre-estimação será quase imperceptível para amostras de grande dimensão, pois haverá mais precisão nas estimativas que estão a estimar as parcelas nulas.

Os dois gráficos da Figura 2 ilustram claramente o que acabámos de referir. As amostras associadas foram construídas a partir do Algoritmo 2, mas, tal como foi dito atrás, resultados similares se obteriam se fossem usados os outros dois algoritmos. Em cada gráfico comparam-se as curvas obtidas para a estatística $T_{\hat{S},n}$ quando fazemos variar o valor de q . Parece claro que a curva com melhor aproximação à normal é aquela em que tomamos $q = m$, seguindo-se a curva em que tomámos $q > m$, com resultados francamente piores quando q é fixado num valor pequeno ($< m$). No gráfico à direita considerou-se um grau de dependência maior ($m = 10$), que como já vimos agrava a aproximação à normal, mas este agravamento foi claramente suplantado com o aumento do tamanho da amostra.

No contexto em que nos situamos, claramente é importante analisar a aproximação à normal nas caudas, de onde são retirados os valores críticos. Na Tabela 1 apresentamos a proporção de valores das estatísticas que caem no intervalo $(-\infty, -1.96] \cup [1.96, +\infty)$. Se a aproximação à normal for razoável, esta proporção deve ser próxima de 0.05. Na tabela estão em comparação as estatísticas T_n , $T_{\sigma,n}$ e $T_{\hat{S},n}$ com diferentes valores de q .

A análise da Tabela 1 vem reforçar as conclusões a que chegámos anteriormente. A Estatística T_σ tem valores muito próximos da estatística T_n para amostras de observações independentes. Quanto à estatística $T_{\hat{S},n}$, a escolha do valor de q afecta bastante os resultados, obtendo-se os melhores resultados para $q = m = 5$ (somando exactamente tantas parcelas quantas as parcelas não nulas da série a estimar). Entre tomar para q um valor grande ($> m$) ou pequeno ($< m$), é notoriamente preferível tomar um valor grande (na tabela $q = 20$).

Passamos agora a analisar o comportamento do estimador \hat{S}_n^2 . Nos gráficos da Figura 3 estão registadas várias estimativas de σ^2 , obtidas com valores

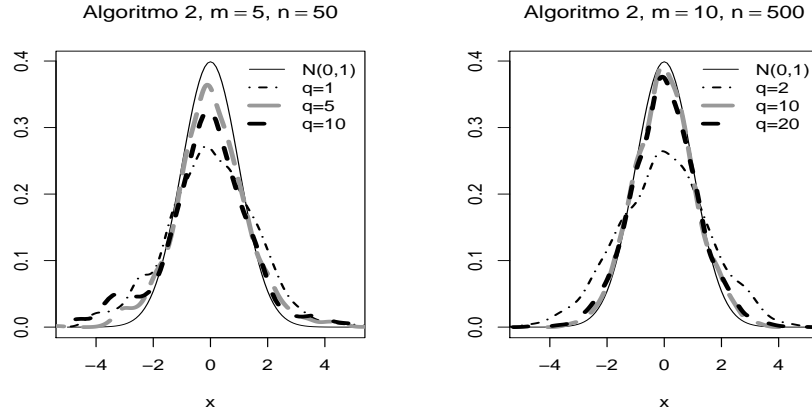


Figura 2: Comparação das estatísticas $T_{\hat{S},n}$ com diferentes valores de q .

Tabela 1: Caudas da distribuição das estatísticas de teste. Amostras associadas geradas pelo Algoritmo 1 com $m = 5$.

n	T_n	$T_{\sigma,n}$	$T_{\hat{S},n}$ com $q = 5$	$T_{\hat{S},n}$ com $q = 1$ ou 2	$T_{\hat{S},n}$ com $q = 20$
50	0.0652	0.0481	0.1272	0.196($q=1$)	0.164
100	0.0558	0.0521	0.0888	0.18($q=1$)	0.1668
500	0.0506	0.048	0.0534	0.1023($q=2$)	0.0871
1000	0.0554	0.0489	0.0518	0.0963($q=2$)	0.0679

diferentes de q . Mais precisamente, cada gráfico tem 20 estimativas de σ^2 (20 bolas) alinhadas verticalmente para cada valor de q . A linha horizontal marca o valor de σ^2 que se pretende estimar. Note-se que foi utilizado o Algoritmo 1 com $m = 20$, portanto, a série a estimar tem apenas $q = m = 20$ parcelas não nulas. Os gráficos revelam um enviesamento bastante acentuado para valores de $q < m$, mas, por outro lado, uma grande variabilidade para valores grandes de q . A solução ideal parece ser $q = m = 20$. É, ainda, claro uma diminuição da variabilidade e do enviesamento quando aumentamos a dimensão amostral. Se recorrermos antes ao Algoritmo 3 para obter amostras associadas, obtêm-se resultados semelhantes. Na verdade, como já foi explicado, uma vez que a série a estimar em σ^2 é convergente, a partir de certa ordem os seus termos são tão próximos de zero que, na prática, tudo funciona como se estes fossem nulos.

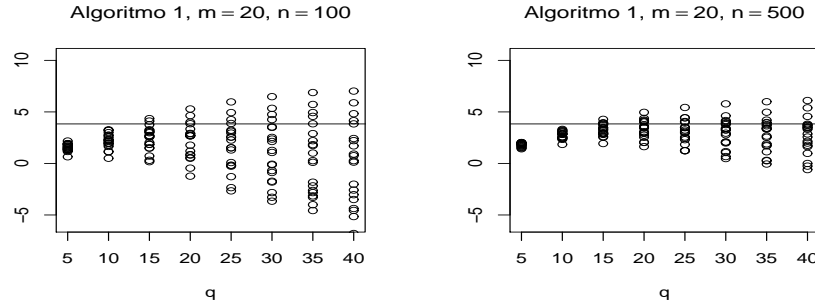


Figura 3: Estimativas da variância σ^2 com diferentes valores de q .

Referências

- [1] Bagai, I. e Prakasa Rao, B.L.S. (1991). Estimation of the survival function for stationary associated processes. *Statist. Probab. Letters*, 12, 385–391.
- [2] Dewan, I. e Prakasa Rao, B.L.S. (2001). *Associated sequences and related inference problems*. Em *Stochastic Processes: Theory and Methods* (D. N. Shanbhag e C. R. Rao, eds.), Handbook of Statistics, Vol 19, 693–731, North-Holland, Amsterdam.
- [3] Dewan, I. e Prakasa Rao, B.L.S. (2005). Wilcoxon-signed rank test for associated sequences. *Statist. Probab. Letters*, 71, 131–142.
- [4] Henriques, C. e Oliveira, P.E. (2003). Estimation of a two-dimensional distribution function under association. *J. Statist. Planning Inf.*, 113 , 137–150.
- [5] Henriques, C. e Oliveira, P. E. (2006). Convergence rates for the estimation of two-dimensional distribution functions under association and estimation of the covariance of the limit empirical process. *Journal of Nonparametric Statistics*, 18, 119–128.
- [6] Henriques, C. e Oliveira, P. E. (2008). Strong convergence rates for the estimation of a covariance operator for associated samples. *Statistical Inference for Stochastic Processes*, 11, 77–91.
- [7] Newman, C.M. (1984). *Asymptotic independence and limit theorems for positively and negatively dependent random variables*. Em *Inequalities in Statistics and Probability* (Y. L. Tong, ed.), IMS Lecture Notes - Monograph Series, Vol. 5, 127–140.
- [8] Roussas, G.G. (1999). *Positive and negative dependence with some statistical applications*. Em *Asymptotics, Nonparametrics and Time Series* (S. Ghosh, ed.) Statist. Textbooks Monogr., 158, 757–788, Dekker, New York.
- [9] Suquet, C. (1994). Introduction à l’association. *Pub. IRMA*, Lille, vol. 34, n° XIII, 1–19.